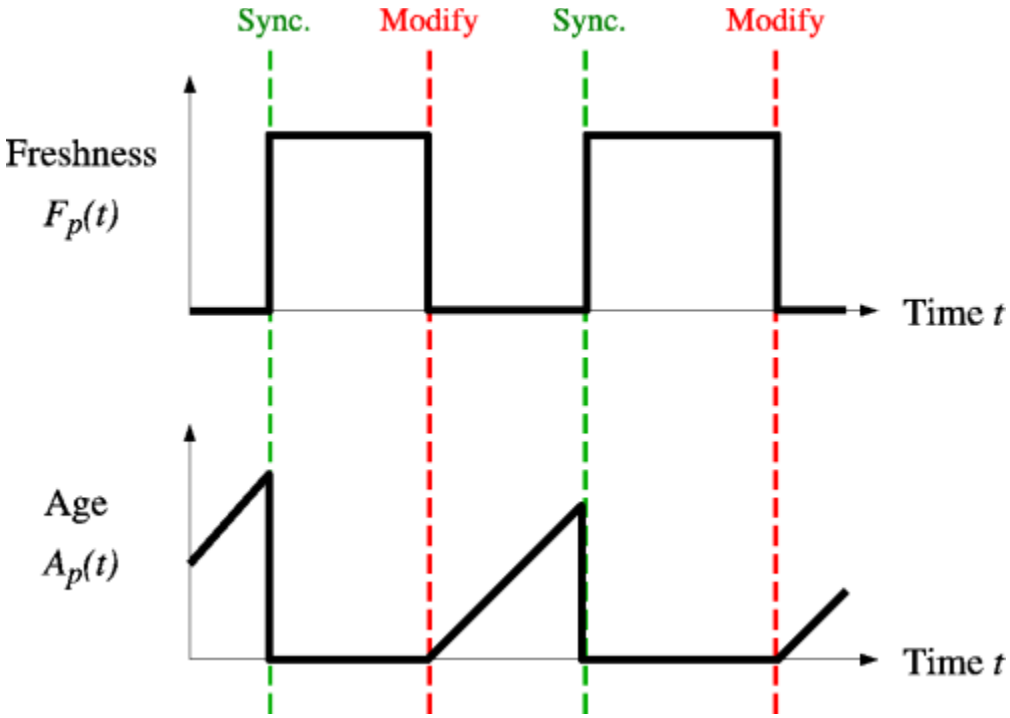
5 Rules For Writing A Web Scraper That Doesn't Break

1. Use a crawling and scraping framework like Scrapy:

Don't try to reinvent the wheel. Frameworks like **Scrapy** abstract many of the complex functions of web crawling like concurrency, rate limiting, handling cookies, extracting links, using file pipelines, handling broken and different encoding to make life easier.

Scrapy also makes life easier by providing the support of selectors for scraping content.



2. Learn & Use XPath or CSS selectors

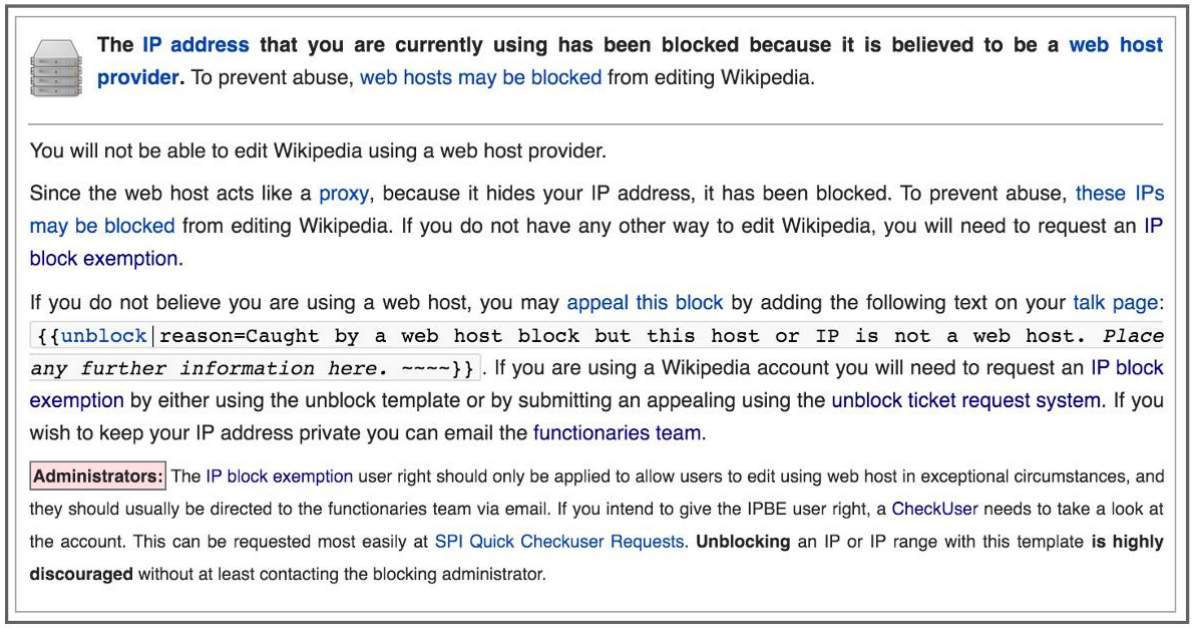
Instead of using RegEx or any other custom rudimentary method to get to the data, you want to scrape, using CSS selectors or XPath or a combination of both makes your code more stable. It protects you against arbitrary changes in a website's code.

3. Scale using Scrapyd and Rotating Proxies

Scrapyd allows you to run multiple spiders at the same time and manage them easily. Combining it with a rotating proxy like **Proxies API** means you can scale your project to dramatic speeds and break a lot of usage and concurrency restrictions of linear coding without incurring the wrath of usage restrictions or IP blocks.

4. Take measures to counter usage restrictions and IP blocks

You might have finally written the perfect scraper that gets every piece of information, managed pagination, code variances, javascript rendering, etc. but it might all come to naught if you get IP blocked. Rotating proxies like **Proxies API** are the way to do overcome it. There is no way around it for serious projects of any decent size, frequency, and importance.



The IP address that you are currently using has been blocked because it is believed to be a web host provider. To prevent abuse, web hosts may be blocked from editing Wikipedia.

You will not be able to edit Wikipedia using a web host provider.

Since the web host acts like a proxy, because it hides your IP address, it has been blocked. To prevent abuse, these IPs may be blocked from editing Wikipedia. If you do not have any other way to edit Wikipedia, you will need to request an IP block exemption.

If you do not believe you are using a web host, you may appeal this block by adding the following text on your talk page: `{{unblock|reason=Caught by a web host block but this host or IP is not a web host. Place any further information here. ~~~~}}`. If you are using a Wikipedia account you will need to request an IP block exemption by either using the unblock template or by submitting an appealing using the unblock ticket request system. If you wish to keep your IP address private you can email the functionaries team.

Administrators: The IP block exemption user right should only be applied to allow users to edit using web host in exceptional circumstances, and they should usually be directed to the functionaries team via email. If you intend to give the IPBE user right, a CheckUser needs to take a look at the account. This can be requested most easily at SPI Quick Checkuser Requests. **Unblocking** an IP or IP range with this template **is highly discouraged** without at least contacting the blocking administrator.

5. Put in checks and balances

There are so many failure points in web crawling projects that you have no control over. It's best to put it a bunch of checks and balances by first identifying them like:

a. Loss of internet connectivity on both ends.

b. Usage restrictions imposed.

c. IP blocks imposed.

d. The target website changes its HTML.

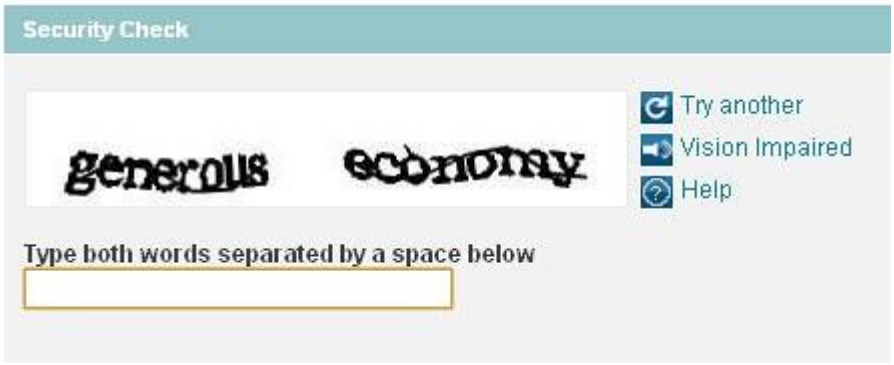
e. The target website is down.

f. Target website issues a CAPTCHA challenge.

g. The target website changes the rules of pagination.

h. The target website uses cookies now.

i. Target website hides content behind javascript.



The author is the founder of **Proxies API**, a proxy rotation API service.